# Project Summary

Mingming Zhang Wenbo Liu Hua Yao

This report focuses on predicting breast cancer by utilizing Support Vector Machine(SVM) and Random Forest Classifier(RFC) machine learning algorithms to analyze the mammography to identify whether benign or malignant mammogram mass.

## Dataset

The dataset is retrieved from UCI Machine Learning Repository named "mammographic\_masses.data.txt" It provides 961 masses data detected in mammography with six numerical attributes: BI-RADS, AGE, SHAPE, MARGIN, DENSITY, and SEVERITY. The SEVERITY attribute is the target to predict. Other attributes are the input attributes for the machine learning models. There are 516 benign cases and 445 malignant cases. During the data preparation, four columns are found with missing values: age, shape, margin and density. Data structure is as follow:



Table

Description automatically generated

To deal with the missing value problem, we will fill a categorical mean into the missing values using fillna() method. This gives us more data for training and testing. However, the values of the features in raw data have different scales. This could affect the correlation between features and the accuracy of the prediction. Therefore, to make each part equally important during the training process, we used a sklearn built-in StandardScaler() function to normalize all the feature values.

## Training and Testing

We used 75% dataset for training and 25% dataset for testing. Therefore, there are 720 cases assigned for training and 241 cases assigned for testing.

Both models are obtained from the built-in model of sklearn library. For SVM, we selected the linear kernel based on the binary classification nature of the problem. For the random forest model, we set the number of trees in the forest to 10 for this simple case. Other parameters are all set to default.

After the models had been trained, we loaded the test data to the models to get the prediction values on the test dataset. To get a more reliable result, we then used the cross-validation method. Finally, having the prediction values, we present the model performance evaluation from confusion metrics and ROC.

## Result

The confusion matrix gives us a breakdown view of prediction results on different classes. We can look into the specific error from the matrix to identify the model problem for further improvement.

Graphical user interface

Description automatically generated with medium confidence Graphical user interface

Description automatically generated with medium confidence

Combining each ROC graph in one figure gives us a clear view of the current best model for mammography prediction. In this case, the simple model SVM did a better job than the complex Random Forest model.

Chart, line chart

Description automatically generated